DOCUMENT RESUME

ED 353 273 TM 018 980

AUTHOR Tatum, Donna Surges
TITLE Controlling for Judge Differences in the Measurement of Public Speaking Ability.
INSTITUTION Chicago Univ., IL. MESA Psychometric Lab.
PUB DATE Apr 92
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Faculty; *College Students; Evaluation Methods; *Evaluators; Higher Education; Individual Differences; *Interrater Reliability; *Measurement Techniques; *Public Speaking; *Speech Evaluation; Speech Skills
IDENTIFIERS Evaluation Standards; FACETS Computer Program; *Rasch Model; Standard Setting

ABSTRACT

Understanding the behavior of those evaluating a speech is important for a complete understanding of the public communication process. Ratings of public speaking were submitted to a Rasch analysis to determine whether objective measurement can create and maintain a standard of speech evaluation. Data used were 1,022 ratings of 168 speeches given by 34 public speaking students at Roosevelt University in Chicago (Illinois) in 2 successive years. Speeches were evaluated by four independent raters and student speakers acting as raters. Raters used an 83-item list of components of a good speech completed by 5 speech teachers. The FACETS computer program of J. M. Linacre provided the means of performing the Rasch analysis. Data do not reveal a pattern to rater severity. Independent raters and teachers were no tougher in their judgments than were speakers. A speaker's ability was not related to how tough he or she was as a rater. Results do indicate that objective measurement can create and maintain a standard of speech evaluation. Three tables and three figures illustrate the study. There is a 14-item list of references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED353273

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

DONNA SURGES TATUM

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Controlling for Judge Differences in the Measurement of Public Speaking Ability

by

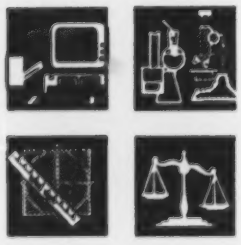
Donna Surges Tatum

MESA Psychometric Laboratory

The University of Chicago

AERA Annual Meeting, 1992

M E S A



TM018980

The University of Chicago

BEST COPY AVAILABLE

Speech Evaluation

Public speaking, by its very nature, produces criticism and evaluation. This commentary is an integral component of the communication process. It is only through feedback that the effectiveness of the communication can be checked. One of the goals of speech evaluation is improvement. Ever since Isocrates, teachers have been trying to enhance their students' natural abilities. Teachers educate them in the theory of public speaking, then make them practice. Eventually most students master the skill, if not the art, of public speaking. A good speech evaluation form is a useful device. It focuses the critique, provides a diagnosis, and guides recommendations for future presentations.

The competencies required for giving a good speech have been studied for centuries. There is commonality to the criteria used when judging a speech. A review of the classical corpus was made to discover evaluative elements of rhetorical criteria. All items were analyzed for redundancy, and a working list was compared with criteria from modern textbooks in the field as well as items included on current evaluation forms to produce the comprehensive list of 83 items covering all aspects of a speech that have been used to judge and evaluate public speaking. Although initially it seemed there were too many items, some similar, or even redundant, a few not well-written, and a tedious format, there also was the confidence that the range of the variable was covered in detail. These items comprise the rating scales used for the evaluation of the speeches in this study.

Five speech teachers sorted the 83 items into their respective rhetorical categories of ethos, pathos, logos and the motivated sequence. There was a great deal of agreement in the selection of the items on the scale used to define each subvariable. After the items were sorted, a discussion was held concerning disagreements. The majority ruled when there was not unanimity on an item's placement.

Data Collection

The data are 1,022 ratings of 168 speeches given by 34 students enrolled in Speech Communication 301 "Persuasive Speaking" at Roosevelt University in two successive years. They will be referred to as "Class 1" and "Class 2". Class 1 was held during an eight-week summer session and met twice a week. Class 2 was a sixteen- week semester course which met once a week.

The characteristics of the sample follow. "ESL" means English as a second language. "-25" represents age of twenty-five or less; "25+" is over twenty-five. "No" means the participant did not have previous training; "Yes" indicates completion of a public speaking class.

TABLE 1. -- Sample Characteristics, Combined Classes, (N =34)

Black = 18	White = 12	ESL = 4
Male = 12		Female = 22
-25 = 12		25+ = 22
No = 18		Yes = 16

The classes were divided into groups of four to eight speakers. Students were assigned five persuasive speeches on topics of their choice. All speeches were taped. Speakers were instructed in the use of the Motivated Sequence, and were to organize their speeches accordingly.

Speakers did not see the rating forms. Guilford (1954) found that raters were easier on everyone if they knew the results would be seen. Raters were informed of the various types of rating errors commonly made (Bock and Bock 1981; Emmert 1989). Guilford felt this awareness was a way to increase accuracy. A Rasch analysis will adjust for these errors and calibrate the nonlinearity of the rating scale.

The independent raters viewed some speeches live and some on tape. The audience and instructor gave verbal feedback and filled out rating forms. The teaching assistants viewed taped speeches privately with the speakers immediately after their presentations. They each rated the speech and discussed ways to improve.

Classroom Evaluation

Kerlinger (1973, 132) defined a rating scale as "... a psychological measuring instrument that requires the rater to assign the rated object to categories or continua that have numerals assigned to them." S. S. Stevens (1968) cautioned that if one would understand the essence of a given measuring procedure, one should ask what was matched to what. One must be sure to have clear definitions and constructs when attempting to match one thing to another. One of the first problems facing the researcher is the selection and application of appropriate criteria. Rating scales are widely used for speech evaluation because they offer a standard set of criteria to be employed for all speeches. They are a systematic way of applying the criteria. It is crucial to provide clear instructions when implementing the rating scale. This helps to achieve matching the presentation, the evaluation of the act, and its numerical description.

The evaluation process is one of the most effective tools a teacher can use. The whole idea of evaluation is to give feedback that can improve the next speech. The speaker as well as the audience gains by the comments. Researchers have explored aspects of speech evaluation in the classroom. Bowers (1964) says that raters should be trained in the use of the scale before using it for the first time. He suggested using a videotaped practice round, followed by a discussion. The classes in this study practiced using the rating form with three live speeches given by the independent raters. This exercise reinforces awareness of what is important in public speaking and is intended to reduce the amount of variation among raters. If each rater uses the rating form consistently, then the Rasch analysis adjusts for the variation in the toughness of the judges.

When done correctly, the students learn that criticism is a useful part of the communication process, and is necessary for improvement. Young (1974) found that students perceive specific comments to be more helpful than general comments. Book and Simmons (1980) advocate balancing the positive and the negative comments. Both types are perceived as helpful. It is important to give enough time between speeches for everyone to complete their evaluations without feeling rushed (Barker, Kibler and Hunter 1968). The instructor must build a positive climate in which to conduct the evaluation (Bock and Bock 1981). All of the above was taken into account and implemented. The groups were small, relaxed and friendly. Thoughtful, kind and honest evaluation was the norm.

Bock and Bock (1981) describe four general types of raters:

1. This type is able to effectively rate both delivery and content; these are the only two factors the rater considers important.
2. This rater does not use any content factor - the only interest is in topic presentation and vocalics.
3. General impressions and verbal adaptations are mostly used in the judgments by those in the third category.
4. Judging tactical matters such as analysis and language is the focus of the last type.

A Rasch analysis will adjust for the different types and toughness of raters as long as they are individually consistent in their use of the rating scale.

Leniency Errors

The rater may be either too easy (positive leniency) or too hard (negative leniency) on all speakers. The Rasch model will adjust for the toughness of the judges, and calibrate them in common units of measure.

Halo Errors

The rater may be too easy (positive halo error) or too hard (negative halo error) on a specific speaker. A Rasch analysis will flag the rater on that speech as "misfitting"

Error of Central Tendency

Many raters have a tendency to group scores around the middle of the scale values. Traditionally, rating scales have five or seven categories. Often the middle category is used as a dumping ground to avoid making a choice. A four or six point rating scale eliminates this problem. McCrosky, Arnold and Pritchard (1967) established that the end points on a semantic differential were further from the points next to them than the other points were from each other. Some raters do not like to make extreme judgments. The Rasch model calibrates the nonlinearity of the rating scale, thus eliminating this concern.

Trait Error

Sometimes the rater has the tendency to be either too easy or too hard on a given item on the rating scale. This is taken into account when the toughness of the raters and the difficulty of the items are calibrated.

Facets Model

For the first time ratings of public speaking were submitted to a Rasch analysis. It is demonstrated that objective measurement can create and maintain a standard of speech evaluation. The data for this research are 1,022 ratings of 168 speeches given by 34 speakers. The FACETS computer program, written by John M. Linacre, provided the means of performing a Rasch analysis.

The model used for this data analysis:

$$\log\left(\frac{P_{nmjgik}}{P_{nmjgik-1}}\right) = B_n - C_j - D_{gi} - F_{gk}$$

B_n $n = 1 - 34$ (speakers)

C_j $j = 1 - 41$ (judges)

D_{gi} $g = 1$, $i = 1 - 83$ (items)

30 items on the Ethos Scale
19 items on the Pathos Scale
29 items on the Logos Scale
5 items on the M S Scale

F_{gk} $g = 1$, $k = 1 - 4$ (4 point rating scale)

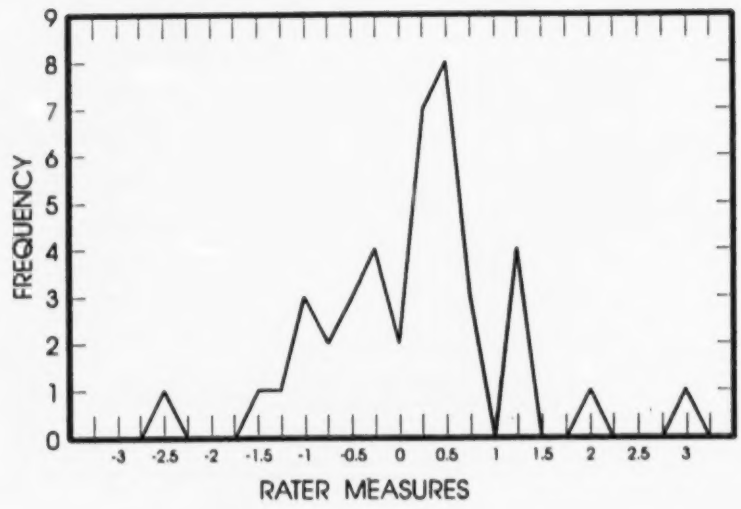
Rater Behavior

Facets computer program allows a researcher to separately analyze various components of the situation under investigation. Raters are an important element of public speaking evaluation. It is necessary to examine their behavior for a complete understanding of the public communication process.

Table 2. --Rater Summary Statistics

Error	Adj. SD	SD MnSq	Infit	SD MnSq	Outfit	Separation	Reliability
.02	.95	0.2	6.1	0.2	5.9	17.58	1.00

Figure 1. --Rater Severity Frequency Distribution



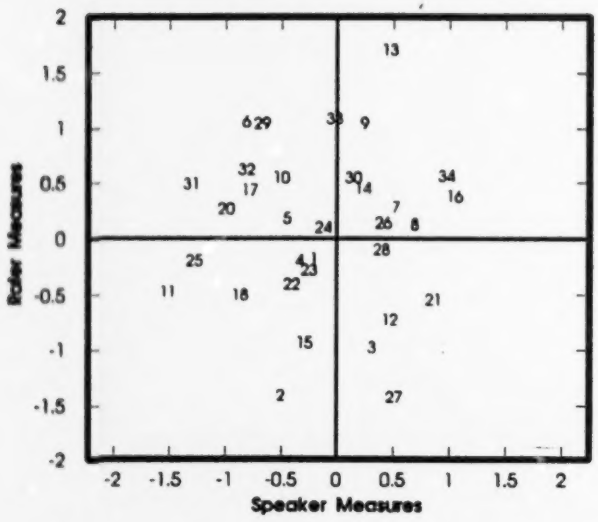
The data show there does not seem to be a pattern to rater severity. An analysis of variance shows that none of the demographic factors was significant. Surprisingly, independent raters and teachers were not any tougher in their judgments than speakers.

Table 3.-- Analysis of Variance of Rater Severity

Squared Multiple R: 0.134

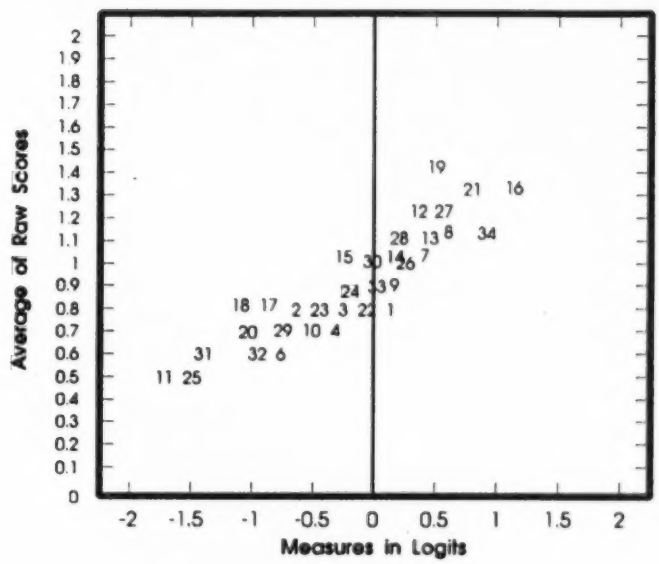
Source	Sum-of-Squares	DE	Mean Square	F Ratio	P
Gender	0.038	1	0.038	0.041	.841
Group	1.265	1	0.633	0.678	.514
Ethnicity	4.290	2	2.145	2.299	.115
Age	0.374	1	0.374	0.401	.531

Figure 2.-- Speaker Ability by Severity as Rater



Again, surprisingly, a speaker's ability is unrelated to how tough she or he is as a rater. Some poor speakers are quite critical, and some good speakers are easy. Evidently a rater's frame of reference and severity is a personal, perceptual experience.

Figure 3.-- Variation of Rater Severity



This graph demonstrates the importance of objective measures rather than a proportion of raw scores. When the toughness of the rater is taken into consideration, the results can be different. For example, eight speakers have a score of .80. However, their measures range from -0.82 to -0.05. Two of the best speakers have scores of 1.3, yet their measures are 1.00 and 0.79. Speaker 19 had the best score of 1.4, but was seventh in ability after scores were conditioned into measures.

Rater Consistency

The four independent raters - 35, 36, 37, 38,- evaluated speeches on tape. Apparently viewing tapes does not affect the severity of the rater. Among all raters, those who viewed tapes range from the toughest to fairly easy raters. However, the independent raters all have slight misfits. This could be due to the fact that some behaviors such as eye contact are difficult to assess on tape. The camera magnifies some nonverbal actions such as facial expressions, or swaying and pacing. These speakers are not trained for the camera. Also, the energy of a live performance is missing. The raters may not have reacted consistently to items on the pathos subscale.

Eight speaker/raters also misfit slightly. This could be due to several reasons. First, perhaps the inconsistent response to the motivated sequence items affects these raters. More time spent on explaining this subscale could help produce more congruity in the raters. Second, the evaluation form used in this study is cumbersome. It is a coupon-book style, and the items are mixed up. Third, the evaluation form is tedious, and rater fatigue may have resulted in inconsistency.

A new and improved evaluation form does correct these problems. Items can be organized according to their respective subscales, making it easier for the rater to maintain a frame of reference. The number of items can be cut down, reducing rater strain.

Selected Bibliography

Arnold, B. Scaling techniques, *Measurement of communication behavior*. Ed Philip Emmert and Larry L. Barker, NewYork: Longman Inc., 1989.

Aristotle, *Rhetoric*. Trans. W. Rhys Roberts. New York: Modern Library, 1954.

Bock, Douglas G., and Bock, E. Hope, *Evaluating classroom speaking*. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills, National Institute of Education Annandale, VA: Speech Communication Association, 1981.

Book, Cassandra, and Simmons, Katrina Wynkoop, Dimensions of perceived helpfulness of student speech criticism. *Communication education*, 1980, 29, 135-146.

Bowers, John W., Training speech raters with films. *Speech Teacher*, 1964, 13, 228-231.

Guilford, J. P., *Psychometric methods*. 2nd ed. New York: McGraw-Hill, 1954.

Kerlinger, Fred N., *Foundations in behavioral research*. 3rd ed. New York: Holt, Reinhardt and Winston, Inc. 1973.

Linacre, John M., *Many-faceted rasch measurement*. Chicago: MESA Press, 1989.

McCrosky J.C. et al Attitude intensity and the neutral point on semantic differential scales. *Public Opinion Quarterly*, 1967, 31, 642-45.

Monroe, Alan H., *Principles and types of speech*. New York: Scott, Foresman, and Co., 1935.

Rasch, Georg, *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press, 1960/80.

Stevens, S. S., Measurement, statistics; and the schemapiric view. Science, 1968, 141, 849-56.

Wright, Benjamin D. and Masters, Geoffery N., *Rating scale analysis*. Chicago: MESA Press, 1982.

Wright, Benjamin D. and Stone, Mark H., *Best test design*. Chicago: MESA Press, 1979.

For further information contact

Donna Surges Tatum
Director of Special Projects

The University of Chicago
MESA Psychometric Laboratory
5835 South Kimbark Avenue
Chicago, Illinois 60637

Telephone: (312) 248-6963
Facsimile: (312) 248-2985